Segmentation of PVA's Donors and Donation Leveraging

Gustavo Brito (r20170760@novaims.unl.pt), Marta Santos (r20170770@novaims.unl.pt),
Miguel Mateus (r20170752@novaims.unl.pt)

Abstract: *PVA's fundraising appeals are one of its strongest assets, especially on the direct mail channel, but optimization and further research are now needed. This study aimed at understanding patterns of the donors' database in order to best channel PVA's efforts. Through the usage of several data cleaning and transformation procedures and unique descriptive learning algorithms, it was possible to develop several robust analysis, compare them and see which one qualified to be the best solution. With the results, an effective and easy-to-implement guidelines for the next marketing fundraising appeal becomes now possible.*

Keywords: Unsupervised Learning, Self Organizing Maps, Segmentation, Nearest Neighbors, K-Means, Python

1. INTRODUCTION

Paralzyed Veterans of America is a NGO that provides programs and services for US veterans with spinal cord injuries or disease. With an in-house database of 13 million donors, PVA is also one of the largest direct mail fundraisers in the United States of America.

The proposed task is to analyze a sample of the results of one of PVA's recent fundraising appeals, containing 95412 donors. This mailing was sent to a total of 3.5 million PVA donors who were on the PVA database. PVA's Mission is to use their expertise to be the leading advocate for:

- Quality health care for their members;
- Research and education addressing spinal cord injury and dysfunction;
- Benefits available as a result of the members' military service;
- Civil rights and opportunities that maximize the independence of their members.

One of the core sources of income for the NGO is the donation on the mailing lists, whether they are made sporadically, monthly, or in a personalized manner. Therefore, the leveraging of such donations becomes crucial, especially in the current times.

The proposition for this study is to perform a Segmentation in such a way that it will be possible for PVA to better understand how their donors behave and identify the different segments of donors/potential donors, analyzing a rich and effective marketing solution afterwards.

2. METHODOLOGY

2.1 Participants

The participants on this database analysis are a sample of 95412 donors, in a total universe of 3.5M PVA's donors. All the participants made at least one prior donation to PVA and they were selected because of such. It is important to denote that although this was the requirement for such collection, there were external data sources considered on this project, namely Metro Mail and Polk. All participant data was collected between September 2013 and June 2017.

The ages of the participants ranged from 1 year-old to 87 years-old and there was no specific division when it came to gender.The mailing includes particular, called Major Donors, that represent the biggest donors to this NGO and therefore their contribution to the cause is of particular interest.

One group that is of particular interest to PVA is *Lapsed* donors. These are individuals who made their last donation to PVA 13 to 24 months ago. They represent an important group to PVA, since the longer someone goes without donating, the less likely they will be to give again. Therefore, recapturing these former donors is a critical aspect that will be analyzed and highlighted throughout this report.

2.2 Materials

The greater helper in this study was Python and its enormous variety of libraries and packages that contained all the statistical methods, functions, and models crucial to this project. It's important to emphasize that some of the most important modelling procedures were made with the usage of the scikit-learn library.

The disclosure of such methods is available on GitHub as an open-source notebook that can be opened here in a variety of notebook environments. Particular emphasis should be given to the developed functions along the report that were of great use during the development of such, and might be useful in some similar analysis.

It is important to highlight that many of the techniques used to make clustering and dimensionality reduction solutions were constructed in forms of functions present on the beggining of mentioned report, mostly adapted from the Data Mining Course for the Master's Degree in Data Science and Advanced Analytics in 2020.

2.3 Procedures

At the beginning of the project, the data was already collected and so the first step was to have a look at such. The research was then conducted in 3 main steps: **data preparation and feature engineering**, the exploration of **clustering methods** and **final adjustments**. The dataset was divided in what was considered its 3 *natural* divisions: personal information about the users, named (Users Data), information about the way customers were donating to PVA (RFA Data) and information on the neighbourhoods the users lived in (Neighbourhood Data).

2.3.1 Data Preparation and Feature Engineering

Users Data

The cleansing of data on this set of variables was made through this guideline:

1. Deleting variables that had a huge proportion of missing values [*]

2. Putting wrongly inputed variables in the right way.

3. Replace *NaN* values with 0, in the cases it made sense to do so, and empty spaces with *NaN*

4. Variables that represent datetime were formatted as such.

In the end, 9 variables were dropped either because they had a lot of missing values('CHILDXX's) they brought no information to the model ('TCODE','OSOURCE'), or their information was already contained in another variable ('GENDER','DOMAIN','MDMAUD'). The variable "MAILCODE" was renamed to "WRONG ADRESS" since it is its true meaning.

Two important analysis were made when it came to the proceeding of the data, the analysis of the 50 states (**The 50 States**) and the analysis of the several variables that represented the income of the person (**Income Analysis**).

The standard deviation was checked for the two variables that represented the donors' wealth to see if the scale between them was too different, which it was, so it wasn't advisable to just join the two variables. Since the variable 'INCOME' had also most of the information needed and 22 % of missing values, While the other two had much more, the two 'WEALTH' variables were dropped.

The next problematic variable was "STATE". The description of the variable indicated that it represents which state a donor belongs to. the variable couldn't have more than 50 unique states, but it had 57. After a brief analysis, it was understood that these extra seven values did not represent US states but USA territories instead. The variable was kept as it was but some renaming of some sort should occur in the future for this variable.

The next step was dealing with the missing values. This was accomplished through imputation methods. On the numeric variables the imputation was done using a KNN imputer with 5 neighbours. The categorical

[*] Also considered non-important or were better represented in other variables.

variables, as they would not be directly used on the clustering solutions, would be imputed in a more advanced stage of this research.

Subsequently, the outlier analysis was performed by using both histograms and boxplots in order to visualize the data and were then removed by filtering the data set.

A coherence check was done on the variable HIT (total number of known times the donor has responded to a mail order offer other than PVA's). If the sum of the known times a donor responded to a mail offer other than PVA's, information present in other variables of the dataset, was larger than the value of HIT, the value would be changed to the sum, thankfully there was no observation with such problem.

To end this step, a correlation analysis was made and if two variables had a correlation bigger than 0.85 one would be dropped, with this in mind only 1 variable was dropped('RECSWEEP').

The feature engineering of the information related to the donors' data consisted of turning the variable DOB to a variable called AGE. For the current date, it was used the last year ever referenced on the dataset (2017).

The information about the gender seemed crucial and the solution found was using the variable 'GENDER'. This variable was transformed in to the dummy variable Female, since originally, it had not only information about the gender, but also if the observation belonged to a joint account.

The variable DOMAIN also contained more information than it should have. This variable contained information about the Urbanicity Level of the Donor's Neighbourhood but also its Socio-Economic Status, with some special requirements.This variable was decomposed accordingly, having, as a result, the variable "Urbanicity" and "SES" (Social Economic Status), that are 2 ordered categorical variables of an ordinal scale.

A variable called All Vets was also created by summing the variables MALEMILI and MALEVET, percentage of males active in the military and male veterans, respectively.

RFA

To treat the variables related to the donor's profile, firstly, all of the variables that represented dates were typecasted into DateTime format. Some basic data inspection was then made, such as seeing whether there were any negative values where they shouldn't exist or which variables had missing values. Most of the date-related variables had missing values, but there was nothing to be done in respect of them as they most likely exist due to a person not receiving a specific promotion or giving a specific gift. There were also some missing values in the variable TIMELAG (number of months between the first and second donation) and NEXTDATE (date of the second donation). It was later found that most of these values corresponded to donors that only donated once and therefore nothing was done as to this, when it was the case that only one donation was made. The variable regarding the date of the first donation (FISTDATE) also had some missing values. Both rows with missing values had a date of a second gift and a difference of months between the first and second donation astronomically high, 1044 and 1088, both rows were dropped.

In the data, there were two variables, FISTDATE and ODATEDW, that supposedly had the same information, the date of the first donation, however, in more than half of the observation, the dates of both columns didn't match. Since all of the DateTime variables would only be used for coherence checks and variable creation, the variable FISTDATE was the only one kept, since it came from the same history file as the other variable regarding the donor's profile.

A series of coherence checks were then made and on the action of such were the following criteria: the severity of the incoherence, the number of observations with such incoherence or the availability of a solution.

Observations with only one donation given but with a value for Timelag (Number of months between first and second gift) were dropped since only 7 rows h this problem.

When the value of LAST DATE (Date associated with the most recent gift) wasn't as recent as the date of one of the gift's given, the value was changed to correctly represent the last known date a donation was given.

0.9% of observations had information that a gift was given when the corresponding promotion hadn't yet been mailed, since this was probably a simple mistake when imputing the dates and the variables were not used in the cluster analysis, the observations were kept.

When the value of the smallest donation given was higher than one of the gifts given or, the value of the highest donation given was smaller than one of the gifts given, the values were changed to represent the respective gifts.

The same procedure was applied to values of the total dollar amount of lifetime donations that were smaller than the sum of the dollar amount all the gifts are given.

Supposedly only lapsed donors were present in the data, however, when comparing the difference between the date that the last promotion was sent to each donor (the "present" date of the data) with the dates corresponding to the gifts given, some of the differences were less than 13 months, meaning the donors were not all lapsed. Since the objective of this study is to understand and cluster based on the lapsed donors behaviours and characteristics, ignoring this discovery would only "dilute" the findings. And so these observations were dropped.

There were also a lot (more than eleven thousand) of observations regarding inactive donors (donors who haven't donated in two or more years). After all the data preparation about 6 % of observations were dropped, these were kept, not only due to the aforementioned situation but also because it allowed us to possibly find some differences between lapsed and inactive donors.

Some more coherence checks were also done, however, no observations appeared to have such problems. All of the observations with a Timelag of zero, and a total amount of donations given higher than one, made the first and second donation on the same day. All of the observations had a total number of promotions received higher than the total number of card promotions received and a date of the first gift older than that of the second gift. All of the observations had a correct Timelag value, meaning a difference in months equal to the difference in months of the first and second gift dates present in the data.

When it came to feature engineering, a few variables were created:

- Total dollar amount of gifts regarding card promotions (types FS, GK, TK, SK, NK, XK, UF, UU).
- Total dollar amount of gifts regarding notepad promotions (types X1, G1).
- Total dollar amount of gifts regarding promotions with only labels or stickers (types LL, WL, CC)
- Recency, in months, regarding the last known gift of each donor.
- Response rate of each donor, the ratio between the lifetime number of gifts to date and lifetime number of promotions received to date.

The average time between gifts (in months) and keeping in mind only the promotions present in the data. When creating this variable, it was noticed that a lot of donors had only given one gift or none, regarding the specific promotions. As such, it was not possible to calculate an average time between gifts for these observations so a value was imputed manually.

It was considered that an imputation such as KNN didn't make particular sense, as the origin of the missing values was known, however the variable seemed to hold a lot of information and the number of missing values was too high to simply eliminate. The value chosen to impute was the highest average time between gifts in the data.

Just like before, outliers were analyzed trough histograms and boxplots of the numeric variables and when two variables had a correlation higher than 0.85, one of them was eliminated. Lifetime number of card promotions and number of card promotions received in the last 12 months were eliminated.

Neighbourhood Data

The first step was analyzing the missing values and only 3 columns had missing values at all, the MSA, the ADI and DMA codes. Upon some research on the meaning of these codes, it was decided that they were not very relevant/meaningful to the analysis, so these columns were deleted.

To analyse the outliers the data set was split into several subsets, this is due to the large number of variables that existed, so by doing this division the analysis can be made more effectively. For this purpose box plots and histograms of the distribution of the data were created and instead of using more conventional methods like IQR, the outliers were removed by the extensive analysis of these graphs and subsequent filtering of the data. Due to the high number of outliers in some variables, it was decided that outliers would only be removed if the problem occured in variable that would be later used either for feature transformation or clustering.

After the outlier analysis and removal, some feature engineering took place since some of the variables could use some transformations.

Firstly the variables IC6 through IC12 were changed to be ICCHEAP, ICMEDIAN and ICEXPENSIVE, this is due to the IC variable containing information about the household income so it can be aggregated into 3 variables. This same aggregation logic was used on 2 other groups of variables, these being the EIC1 through EIC14 which were turned into 1stSECTOR, 2ndSECTOR and 3rdSECTOR (the EIC variable represented employment industry so it was just replaced with its respective work sector), and to the variables EC2 through EC8 which were turned into 'Less Educated', 'High school' and 'University' (again since these variables represented education, it was correct to aggregate them like this to reduce the number of dimensions).

Even after these transformations, the dimensionality of the data set was still a problem. The data was normalized through z-score standardization and the dimensionality-reduction approach was the one to follow. We decided to proceed with a Principal Components Analysis (PCA) applied to reduce the number of dimensions, as such a division of perspectives of the PCA analysis was conducted. The final result were 4 PCA Analysis done in the following perspectives: **PCA Ethnicity, PCA Family Neighbourhood, PCA Housing Conditions, PCA Development**. The correlations of the PC with its variables can be found in the end of this analysis.

The number of principal components (PC) chosen using this technique was done by using both the scree plot and the cumulative variance plot, the number of PC would then be not only the ones represented on the elbow of the scree plot, but that also accounted for at least 80 %of the total variance.

During the Principal Components Analysis, the criterion to better understand how suited the data, more specifically the different chosen variables, is for the PCA used is the Kaiser-Meyer-Olkin (KMO) Test. The test measures sampling adequacy for each variable in the model and for the complete model. The statistic is a measure of the proportion of variance among variables that might be common variance. The lower the proportion, the more suited your data is to perform a PCA. KMO returns values between 0 and 1. A rule of thumb for interpreting the statistic is: *If the value is between 0.8 and 1 the sampling is adequate.* and *If the value is less than 0.6, the sample might not be adequate and that remedial action should be taken.*

A more interesting view on interpretation of the statistics by the author himself and a detailed description of the hypothesis test can be found here. The KMO Scores for the 4 Principal Component Analysis were as follows:

PC Perspective	KMO Score
Ethnicity	0.686
Family Neighbourhood	0.879
Housing Conditions	0.807
Development	0.912

Table 1: KMO Scores for the PCA Analysis

PCA Ethnicity

The first perspective tackled was the ethnicity, this was accomplished by using the variables that related to this topic ETH1,2 and 5, ETHC1 through ETHC6, LSC1 and 2 and POBC1 and 2. After analysing the scree plot and the accumulated variance plot 3 principal components were retained.

The 1st PC has high correlations with variables relating to black communities so it was named accordingly 'Black Communities' , the second principal component was mainly correlated with Hispanic communities and again was named 'Hispanic communities' the third and final principal component was only correlated with variables related to white kids born in the state of residence so it was named 'Native white kids'.

PCA Familiy Neighbourhood

The second perspective found was named Family neighbourhood, as the variables chosen were related to the family and accommodation situation in a neighbourhood (i.e. number of residents in a house, number of children in a residence so on and so forth). With the same criteria as before, 3 PC were kept.

The 1st PC had high correlations with variables pertaining mainly to neighbourhoods with unmarried adults with a low number of people per housing unit and a low amount of children, this earned it the name of 'Single Adults Neighbourhood'.The 2nd PC was somewhat similar to the first one but this time had high correlations with variables relating to children so it was named 'Single Parents Neighbourhood'. The 3rd PC was similar to the first one with a low amount of people per room unmarried adults and no children, but this time there was a difference in the age variables, indicating a younger age, so it was named 'Younger Neighbourhood'.

PCA Housing Conditions

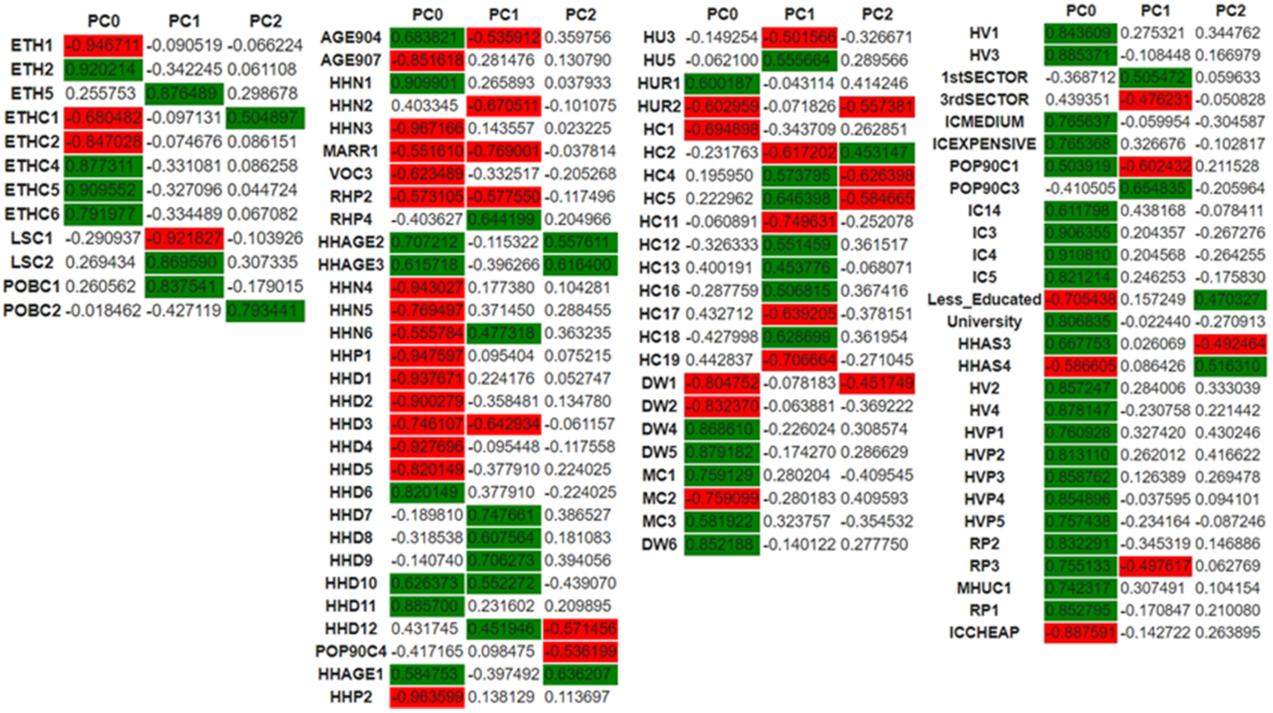
The 3rd perspective analyzed variables relating to the housing conditions of the different neighbourhoods, meaning the pluming conditions access to electricity, heating system, along with others. 3 principal components were kept, in order to maintain at least 80% of the variance.

The 1st PC had high correlations with variables like central plumbing a low room count and in high unit structures (buildings) so it was labelled 'Modern Flats'. The 2nd PC revealed very different results from the first since most of the highly correlated variables, referred to low number of years since the houses have been built,and also that these houses were in less developed areas. As most of the variables pointed to the date that the houses were built in this principal component was named 'New Houses'. The 3rd PC indicated a higher value in the number of rooms than the first component and a fairly high number of years of occupation so it was named 'Old Dwelling'.

PCA Development

For the fourth and final perspective, the variables relating to the level of development of the neighbourhoods were chosen, utilizing variables that would reflect that such as, level of education, the value of the houses themselves, the average salaries of the residents and the like. 3 principal components were retained.

The 1st PC had a lot of incredibly high correlations, most of them on variables that pointed to wealthier areas like a very high education level, high average salaries, high house pricing, so it was named 'Wealthier Areas'.The 2nd PC was, in contrast, to the first one with low levels of education accompanied by population outside urbanized areas and mostly working in the first activity sector, so it was named 'Rural Areas'. The third and final principal component was a bit tougher since it had high correlations with a low level of education and people living at about the poverty level, but the house prices were not very low. This was solved by naming it 'Gentrified Areas' since gentrified areas tend to be areas with a population living in bad or run-down conditions but the city hall decides to improve this area by renovating it but not improving the salaries and living conditions of the residents, which would explain the average house prices but low level of education and very low wages.



	PC0	PC1	PC2
ETH1	-0.946711	-0.090519	-0.066224
ETH2	0.920214	-0.342245	0.061108
ETH5	0.255753	0.876489	0.298678
ETHC1	-0.680482	-0.097131	0.504897
ETHC2	-0.847028	-0.074676	0.086151
ETHC4	0.877311	-0.331081	0.086258
ETHC5	0.909552	-0.327096	0.044724
ETHC6	0.791977	-0.334489	0.067082
LSC1	-0.290937	-0.921827	-0.103926
LSC2	0.269434	0.869590	0.307335
POBC1	0.260562	0.837541	-0.179015
POBC2	-0.018462	-0.427119	0.793441

	PC0	PC1	PC2
AGE904	0.683821	-0.535912	0.359756
AGE907	-0.851618	0.281476	0.130790
HHN1	0.909901	0.265893	0.037933
HHN2	0.403345	-0.670511	-0.101075
HHN3	-0.967166	0.143557	0.023225
MARR1	-0.551610	-0.769001	-0.037814
VOC3	-0.623489	-0.332517	-0.205268
RHP2	-0.573105	-0.577550	-0.117496
RHP4	-0.403627	0.644199	0.204966
HHAGE2	0.707212	-0.115322	0.557611
HHAGE3	0.615718	-0.396266	0.616400
HHN4	-0.943027	0.177380	0.104281
HHN5	-0.769497	0.371450	0.288455
HHN6	-0.555784	0.477318	0.363235
HHP1	-0.947597	0.095404	0.075215
HHD1	-0.937671	0.224176	0.052747
HHD2	-0.900279	-0.358481	0.134780
HHD3	-0.746107	-0.642934	-0.061157
HHD4	-0.927696	-0.095448	-0.117558
HHD5	-0.820149	-0.377910	0.224025
HHD6	0.820149	0.377910	-0.224025
HHD7	-0.189810	0.747661	0.386527
HHD8	-0.318538	0.607564	0.181083
HHD9	-0.140740	0.706273	0.394056
HHD10	0.626373	0.552272	-0.439070
HHD11	0.885700	0.231602	0.209895
HHD12	0.431745	0.451946	-0.571456
POP90C4	-0.417165	0.098475	-0.536199
HHAGE1	0.584753	-0.397492	0.636207
HHP2	-0.963599	0.138129	0.113697

	PC0	PC1	PC2
HU3	-0.149254	-0.501566	-0.326671
HU5	-0.062100	0.555664	0.289566
HUR1	0.600187	-0.043114	0.414246
HUR2	-0.602959	-0.071826	-0.557381
HC1	-0.694898	-0.343709	0.262851
HC2	-0.231763	-0.617202	0.453147
HC4	0.195950	0.573795	-0.626398
HC5	0.222962	0.646398	-0.584665
HC11	-0.060891	-0.749631	-0.252078
HC12	-0.326333	0.551459	0.361517
HC13	0.400191	0.453776	-0.068071
HC16	-0.287759	0.506815	0.367416
HC17	0.432712	-0.639205	-0.378151
HC18	-0.427998	0.628699	0.361954
HC19	0.442837	-0.706664	-0.271045
DW1	-0.804752	-0.078183	-0.451749
DW2	-0.832370	-0.063881	-0.369222
DW4	0.868610	-0.226024	0.308574
DW5	0.879182	-0.174270	0.286629
MC1	0.759129	0.280204	-0.409545
MC2	-0.759099	-0.280183	0.409593
MC3	0.581922	0.323757	-0.354532
DW6	0.852188	-0.140122	0.277750

	PC0	PC1	PC2
HV1	0.843609	0.275321	0.344762
HV3	0.885371	-0.108448	0.166979
1stSECTOR	-0.368712	0.505472	0.059633
3rdSECTOR	0.439351	-0.476231	-0.050828
ICMEDIUM	0.765637	-0.059954	-0.304587
ICEXPENSIVE	0.765368	0.326676	-0.102817
POP90C1	0.503919	-0.602432	0.211528
POP90C3	-0.410505	0.654835	-0.205964
IC14	0.611798	0.438168	-0.078411
IC3	0.906355	0.204357	-0.267276
IC4	0.910810	0.204568	-0.264255
IC5	0.821214	0.246253	-0.175830
Less_Educated	-0.705438	0.157249	0.470327
University	0.806835	-0.022440	-0.270913
HHAS3	0.667753	0.026069	-0.492464
HHAS4	-0.586605	0.086426	0.516310
HV2	0.857247	0.284006	0.333039
HV4	0.878147	-0.230758	0.221442
HVP1	0.760928	0.327420	0.430246
HVP2	0.813110	0.262012	0.416622
HVP3	0.858762	0.126389	0.269478
HVP4	0.854896	-0.037595	0.094101
HVP5	0.757438	-0.234164	-0.087246
RP2	0.832291	-0.345319	0.146886
RP3	0.755133	-0.497617	0.062769
MHUC1	0.742317	0.307491	0.104154
RP1	0.852795	-0.170847	0.210080
ICCHEAP	-0.887591	-0.142722	0.263895

Figure 1: Correlations of the PC with its variables

After the PCA analysis was concluded, a correlation analysis was made to the neighbourhood data set with the same criterion as before. It's important to denote that priority was given to the previously retrieved principal components.

Just before the decision on how to proceed with the unsupervised learning method, it was decided that the information about the users should be united with the data referencing the neighbourhood. As so, the information about the Veterans and Military Staff that live in the surrounding areas of the donors was the one that prevailed lastly. All of the numeric variables that were not yet standardized also went through the z-score normalization.

2.3.2 Clustering Methods

In the final step of this solution, two perspectives were chosen for clustering. The first one being the donor's profile in regard to how they donate to PVA and the second one the demographic characteristics of the donors, more specifically the characteristics of their neighbourhood. As such two different clustering analysis were conducted and then consolidated into one.

An alternative to the following clustering methods considered was **A Priori Data Segmentation Strategies**, more specifically a RFM Analysis for the donor profile variables. This analysis can be made through two perspectives: By quartiles/quintiles or Hard Coding. The procedure chosen was quartiles.

The quartiles were divided and there were now 2 alternatives to follow. The first one consists on concatenating the quartiles' scores per observation, e.g, a person that had belonged to the 2nd quartile in recency, third in frequency and fourth in monetary, would be labeled 234. The second alternative was to sum the quartiles' scores, in this specific case, this person would get a score of 2+3+4 = 9.

This approach was ignored in the end since a more complex cluster analysis, with more variables, would contains more information about the donors and about their consumption habits, which will be useful in the marketing campaigns.

The variables for each cluster solution were chosen initially on a subjective criteria, based on their perceived importance for the problem at hand and with constant adaptation based on the gathered results. But a more mathematical approach was also taken, and the Hopkins Test Statistic was used to assess the "clustering tendency" of a set of variables, *a priori* of a clustering analysis.

The **Hopkins statistic** is used to assess the clustering tendency of a data set by measuring the probability that a given data set is generated by a uniform data distribution. In other words, it tests the spatial randomness of the data. The Hopkins statistic can be calculated as follow:

$$H = \frac{\sum_{i=1}^{m} u_i^d}{\sum_{i=1}^{m} u_i^d + \sum_{i=1}^{m} w_i^d}$$

Values 0.5 indicate that the data is regularly spaced and 0.5 means the data is random. If the value is between 0.7, ..., 0.99, it has a high tendency to cluster.

The variables chosen for the analysis of the donor's profile were: Average Time Between Gifts, Response Rate, Recency, Frequency, Monetary and Average Gift, with a Hopkins statistic of 0.98.

The variables chosen for the demographic perspective clustering were: Black Communities, Single Adult Neighbourhood, Modern Flats, Wealthier Areas, Single Parent Neighbourhood, New Houses and All Vets with a Hopkins statistic of 0.93.

A total of four different clustering methods were considered for each perspective: K-Means Algorithm, Hierarchical Clustering with K-Means centroids, K-means with SOM units and Hierarchical Clustering on top of SOM units. Due to the high computational costs required, unfortunately it was not possible to perform Hierarchical Clustering with the data.

Hierarchical Clustering on top of K-means, using the k-means algorithm first, producing a lot of clusters(around 200) and then applying the Hierarchical Clustering on the centroids produced by this k-means and assigning the corresponding cluster label to each observation.

K-means clustering on top of SOM units. Much like the previously explained method, firstly an algorithm was used to find the correct number of clusters. This time utilizing SOM with a sizable grid (2500 units), this is the intended use of SOM since it initializes by projecting the grid onto the data, then fitting this grid to the data at hand and returning the component plates, the u-matrix and the hit-map. By observing the u-matrix, one can extract the number of natural groups of data. The observation of the u-matrix is then followed by running the k-means according to the number of clusters that were found.

Hierarchical Clustering on top of SOM, this, much like the second model was accomplished by running a SOM algorithm first which works identically to the first utilization of the SOM algorithm , getting the number of clusters and then using the Hierarchical Clustering to join the units produced by the SOM algorithm. This is accomplished by again utilizing the Euclidean distance between the Best Matching Units and clustering the ones that have the least distance until there are only the number of cluster specified by the user.

The main objective of making several perspectives is that the results can vary from method to method and there is no logic way to guarantee *a priori* that one will consistently perform better than the other. The methodology was, therefore, trying the methods that would best fit the data, comparing results, and keeping the method that provided the most insights.

To decide on the number of clusters for each perspective, an elbow plot of the Kmeans algorithm inertia and average silhouette score, the dendrogram of the different Hierarchical solutions and the SOM units U-Matrix were analyzed.

Most of them seemed to indicate 4 clusters and so that was the number chosen, which also allowed the comparison of R-squares of the different clustering methods. However 3 and 5 clusters were also tried. Three clusters would always group more than half of the observations in a single cluster and five gave very similar results to four clusters, so the "simpler" solution was kept.

The R^2 for the 4 methods of clustering in both perspectives were as follows:

Donor Profile Perspective	R^2
K-Means	0.50
Hierarchical+K-Means	0.29
Hierarchical+SOM	0.46
K-Means+SOM	0.49

Demographic Perspective	R^2
K-Means	0.40
Hierarchical+K-Means	0.30
Hierarchical+SOM	0.33
K-Means+SOM	0.38

Tables 2 and 3: R-squared of the different cluster solutions

The chosen algorithm for both perspectives was the K-means algorithm, as it provided the best scores (R^2), and the different algorithms gave very similar clustering interpretations.

After having both cluster solutions all that was left to do was to merge them. Naturally, merging two cluster solutions with 4 clusters each provided a final number of clusters of 16. As expected, all of this "final" clusters have very few observations each and there is no need to proceed with 16 clusters and 16 subsequent different marketing campaigns. To combine some of this final clusters a Hierarchical Algorithm was used on the cluster centroids and after analyzing the corresponding dendrogram and some trial and error it was concluded that the optimal final number of clusters was also 4.

3. RESULTS

3.1 Final Clustering

- **Cluster 1:** Individuals that live in non-black communities with family-like households that have higher levels of education and income. The houses in these neighbourhoods are usually new dwellings with prices above average, characterized by its modern conditions. On the donor profile segment, they behave normally.

- **Cluster 2:** Individuals that have better engagement with PVA's promotions and better performance on RFM indicators, although, on average, their donations are smaller per promotion.

- **Cluster 3:** Individuals that have donated consistently and recently (having in to account that only lapsed and inactive donors are being analysed), although they do it less frequently and their donations have lower values.

- **Cluster 4:** Individuals that haven't donated frequently and have less engagement with PVA, but donate more per promotion.

Clusters 2,3 and 4 behave similarly on the demographic clustering variables. They are characterized by above average presence of Black Communities and small households, in poorer neighbourhoods with new modern flats. These are not neighbourhoods in which typically military staff or veterans prevail.

Cluster	Nr of Observations
1	32478
2	19156
3	25988
4	12284

Table 4: Number of Donors per Cluster

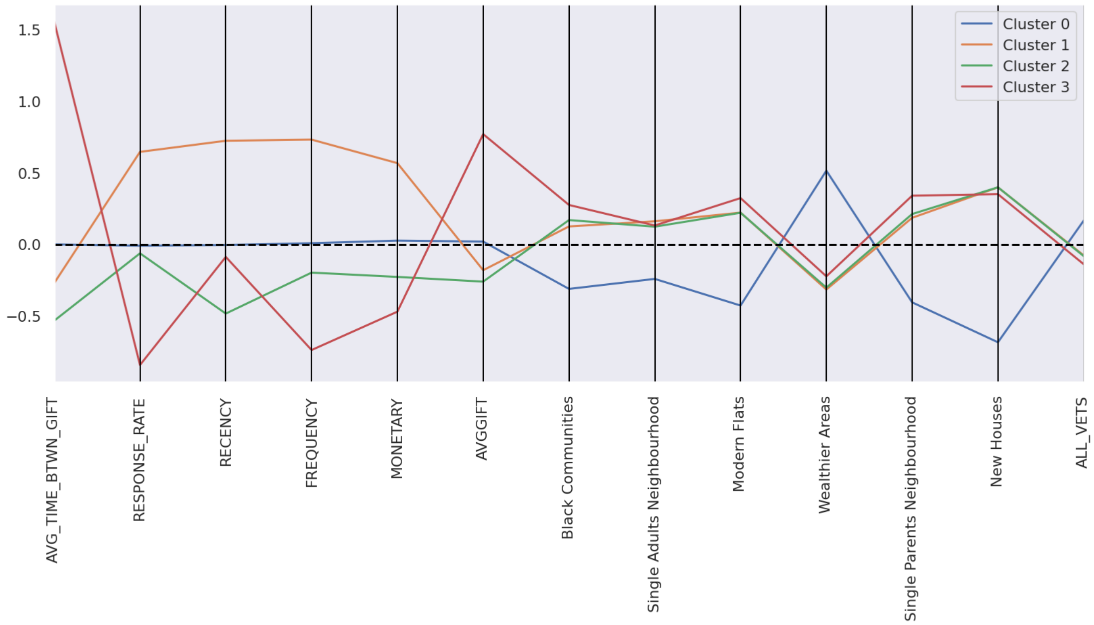


Figure 2: Average value of the Chosen Variables per Cluster

The Distribution of Values of each variable per cluster can be found in here

3.2 Final Adjustments

3.2.1 Categorical Variables Analysis

To analyze some of the categorical variables present in the data, to better characterize each cluster, some missing values had to first be imputed and predictive models were used to full fill that task. Deleting the rows with missing values would be sub-optimal, since some of the observations used to produce the clustering analysis would be lost, and some of the variables had a number of missing values too high for a simple mode imputer.

The method chosen for imputation was Gradient Boosting, a technique that uses several weaker classifiers iteratively, more specifically decision trees, to reach a more robust classifier, through improving the errors of each prior weaker classifier constantly.

To impute missing values, several new data sets were created, one with the rows that did not have the missing values and the columns used in the final clustering solution as independent variables, which would be used in the train and validation split, and another with the rows corresponding to the rows with missing values, and the same columns as the first data set. These data sets were replicated every 4 times since there were 4 distinct target variables with missing values in different rows. The following scores were attained:

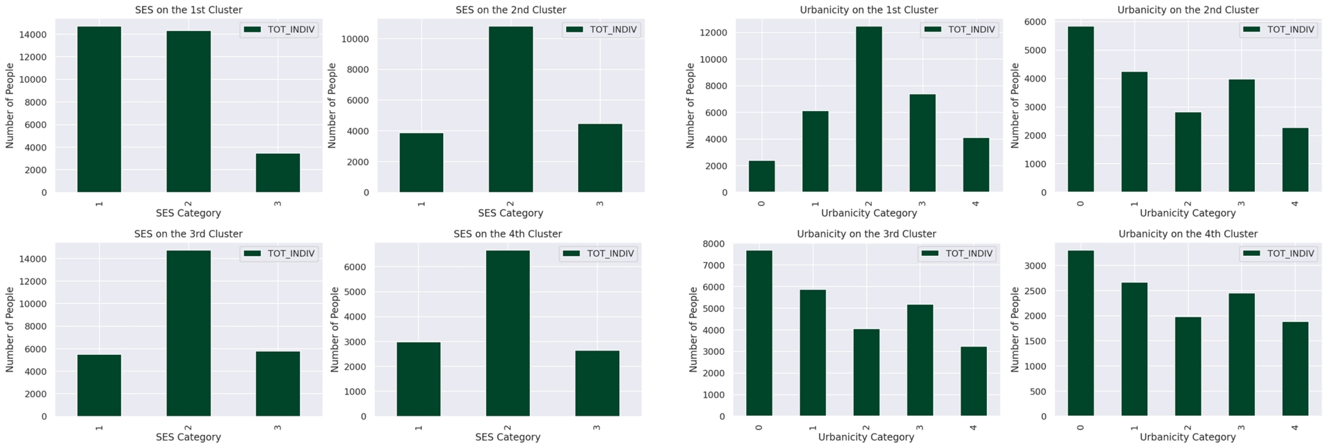
Variable	Train Score	Test Score
SES	0.798	0.761
URBANICITY	0.668	0.610
INCOME	0.669	0.274
DATASRCE	0.625	0.606

Table 5: Scores on the Imputation of Missing Values

The models were then used to predict their respective missing values. [†]. The analysis were made divided by cluster, so that the number of individuals in each cluster did not influence it.

The **socio-economic status** [Figure 4] of the neighborhood the individual lives in reflects their income and social capabilities. On the raw data it was reflected a difference on SES between urban communities and the rest, however, in previous data transformations this disparity was eliminated. As so, and accordingly to the analysis made on the clusters, it is seen that on the 1st cluster the individuals live, in average, in wealthier areas, having 1 (Highest SES) as the most common label. Also according to the denominations of the clusters, the 2nd, 3rd and 4th clusters have a worse performance on this indicators, but don't show much disparity among them when it comes to the information of socio-economic status.

The **Urbanicity** [Figure 5] of the clusters was transformed in an ordinal variable, that had, originally the following values: 1=Urban, 2=City, 3=Suburban, 4=Town, 5=Rural. As a big overview of the graphs, it is possible to see that the 1st cluster is the one that is distinguishable from the rest. As such, on the 1st cluster it is possible to see that the majority of individuals live in suburban areas, which reinforces the idea behind the tendencially wealthier zones with houses with larger areas, as represented in its cluster interpretations. On the 1st cluster the percentage of individuals that live in towns and rural areas is considerably higher than on more urbanized areas. Clusters 2,3 and 4 behave in similar ways, having higher values considering the urbanized areas, and less evidence of living in more rural areas, although the 4th cluster shows less diversity on this variable.



Figures 4 and 5: SES Level per Cluster[4] and Urbanicity Levels per Cluster [5]

The **DATASRCE** [Figure 6] contains information on the source of overlay data for the donors. 1=MetroMail, 2=Polk, 3=Both. This information is useful as it will be possible for PVA to understand which might be the best mailing solution for each cluster in the analysis. Once again, it is observable that the 1st cluster behaves significantly different from the rest of the clusters. In the 2nd, 3rd and 4th clusters there is predominantly a preference for both mailing solutions, however, the proportion of Polk only is also important. On the other hand, the 1st cluster has shown a big preference for having both mailing solutions, and the disparity between MetroMail and Polk is not as colossal as the previous referenced ones. The recommendation on this matter is that the usage of both companies might be a good option, but in case the preference of one should be made, it should be Polk as it seems to be more consolidated in the market.

[†] The low scores on the prediction of INCOME, and with correspondent higher number of missing values, did not provide enough confidence on its results to follow through with this variable.

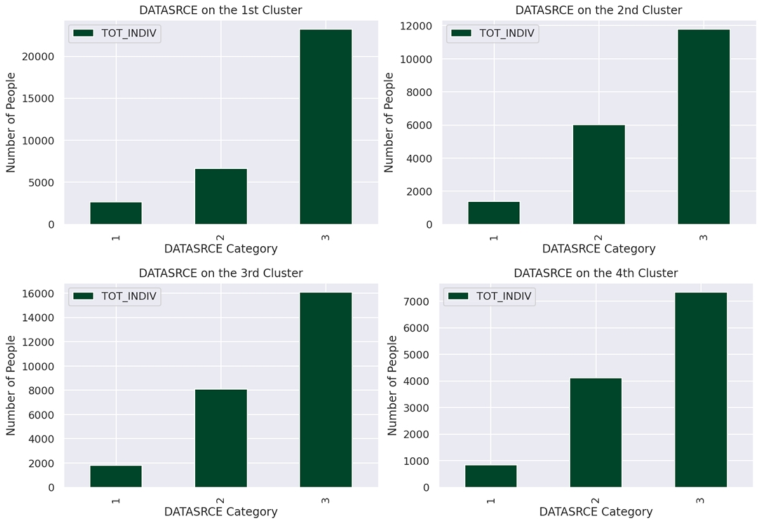


Figure 6: External Data Source per Cluster

The following categorical variables considered to be analyzed were the ones that represented the knowledge on the **interests of the individuals**. The graphics per cluster and per interest can be found here. Since now direct comparisons between clusters are being made, the proportion between clusters was analysed. An example is as follows:

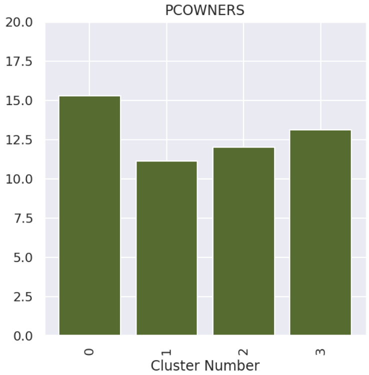


Figure 7: Proportion of Individuals w/ Interest on PC

The distribution of interests is fairly equal among all clusters, however, on most of the 18 interests, it is possible to see that there is a slightly higher value of interests on the first cluster. To assess this, an average of interests per person per cluster was made. Its important to highlight that this only qualifies whether an individual has interest on the topic or not, not quantifying it. As it was believed that the quantification was necessary, the same analysis was made on the information on known times the donor has responded to similar interests. [‡]

Both analysis provided the following results:

[‡] the number of missing values on both perspectives was relatively high, however, the proportion of missing values per cluster was fairly equal and having always in to account that the information relates to the known interests, meaning a zero or a missing value simply means that it is not known whether the donor has that interest or not, comparisons can be made

Cluster	Interest (Qualification)
1	1.48
2	1.29
3	1.42
4	1.28

Variable	Interest (Quantification)
1	2.21
2	1.54
3	1.49
4	1.32

Tables 6 and 7: Average number of interests per person

This tables show that on the 1st cluster there is a clear prominence on the purchase of cultural products, both in qualification and on quantification. When the quantification is introduced, the difference between the 1st cluster and the rest becomes higher. This information might tell us (and remembering once again that the 1st cluster represent people who tendencially live in wealthier areas) that donors that belong on the 1st cluster are more likely to purchase cultural products and, as so, this might be a good hint for the marketing campaigns. A prioritization could be made also on the 3rd and 2nd clusters to apply this information.

3.2.2 Re-Introduction of Observations

As previously mentioned, not all observations were used during the cluster analysis. Having the possibility to estimate a predictive model to classify the removed observations, there was no reason not to reintroduce them in the cluster solution and also making them a part of the different marketing campaigns.

A predictive model was therefore chosen, more specifically an AdaBoost Classifier. AdaBoost classifier is a meta-estimator that initially fits a classifier to the data, a decision tree composed of only one split, and subsequently fits copies of the initial classifier on the same data but giving more weight to observations that were previously wrongly classified. AdaBoost's hyper parameters were estimated with a grid search with a cross validation of 3. Typically, it is best to have a higher number of partitions in the cross validation, especially due to the abundance of observations to use for training, however, it was exactly this fact that made it not viable to use more partitions due to the high computational cost and time required to train the models.

The independent variables used were the exact same ones used for clustering, and the dependent one was naturally the cluster labels. The observations used for clustering were divided in to two partitions, train (70%) and validation (30%), having the dependent variable stratified.

After estimating the hyper parameters, the corresponding model had an accuracy of 0.7 for both train and validation, and so, the observations were classified accordingly.

861 observations were added to the first cluster, 1091 to the second, 975 to the third and 2579 to the fourth.

With a predictive model estimated, and having in to account that the chosen predictor has in its base architecture decision trees, it was possible to asses the feature importance of each variable to the final prediction through the mean decrease in impurity provided by each feature, giving us a sense of which variables have a higher impact when classifying new observations.

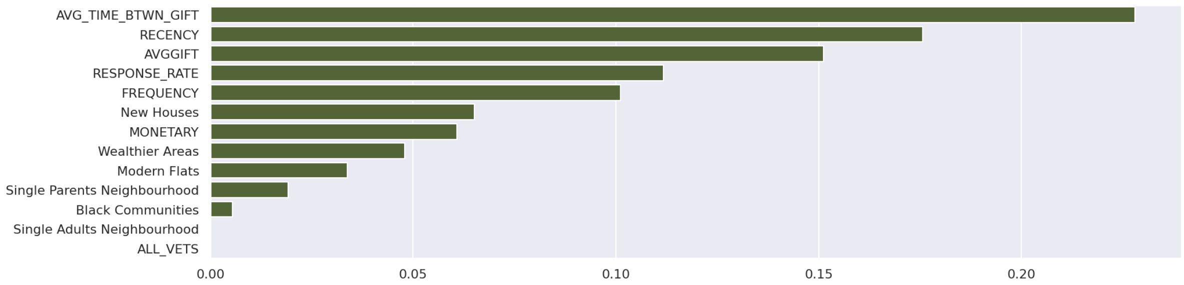


Figure 8: Feature Importance of the Chosen Variables

4. DISCUSSION

4.1 Marketing Guidelines

As shown throughout the report, the marketing campaigns was always kept in mind and now comes the time for its explanation. It's important to highlight that the following guidelines are only that, meaning, the universe of solutions for this clustering analysis is immense.

- **Cluster 1:** As the donors on this cluster are people who live in wealthier neighbourhoods and seem to have a better engagement with mailing offers (namely many that relate to hobbies or cultural subjects), it is believed they are potential customers, even though they do not interact neither above or below average with PVA's promotions. Therefore, to attract their attention, a marketing campaign should be developed accordingly to their personal interests, of the type by X product and a percentage of the purchase goes towards PVA. The offer should relate to the interests that the individual has shown (present in the data), and for those in which there is no information, a more generic approach should be taken, accordingly to the cluster's most common interests, as stated on Section 3.2.2.

- **Cluster 2:** The approach considered best to this cluster constitutes on a give-away, as this would make these previously engaging and valuable donors more prone to donate rapidly again. An example that could be made is proposing a donation of 10$ for a chance to win a bicycle as follows:

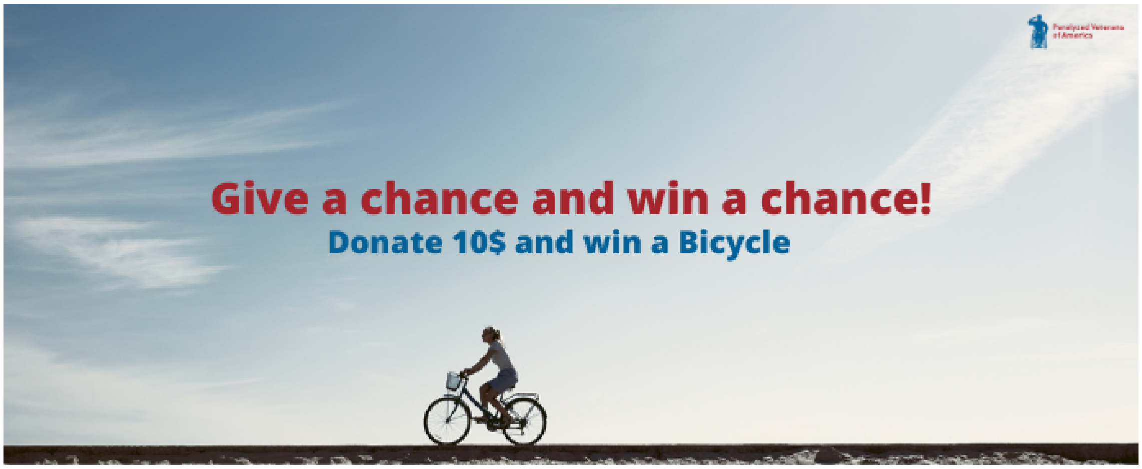

Paralyzed Veterans of America
Give a chance and win a chance!
Donate 10$ and win a Bicycle

Figure 9: E.g of a Marketing Campaign for the 2nd Cluster

- **Cluster 3:** For this cluster the main objective is to raise awareness on the Paralyzed Veterans topic. As the individuals in this cluster tendencially have less engagement and make smaller donations, understanding the true cause of the NGO would be crucial to their reinsertion on the cause. As so, several free options are available, and according to PVA's website, the organization has Research Foundations and several partnerships that would be of interest to the Lapsed Donors. One potential idea is offering a visit to the Spinal Cord Research Labs and several other museums located on the persons' region.

- **Cluster 4:** The donors of this cluster will be introduced to a Reward System. As these donors only donated a few times before they stopped their donations a reward system should be put in place to retain them. This system should give rewards not only for more donations but also according to donations of a higher monetary value, since the monetary values of their donations are above average.

4.2 Limitations

The 1st considerable limitation was the lack of previous treatment on the data. This lead to a lot of time consuming data cleaning and preparation, as well as feature transformation and dimensionality reduction. Another constraint was the amount of missing values in the existing data, which didn't allow the usage of several variables.The 3rd limitation was the lack of computing power, as some analysis like the Hierarchical Clustering was impossible to be made.

5. CONCLUSIONS

The main objective of this study was to segment the customer database to extract value from PVA donors' and understand how this intrinsic knowledge could benefit the NGO and the ones interested in donating. The final step of this analysis was therefore a sketch of what the ideal direct mail marketing campaign would be like with the information from the donors.

The clustering of the final assessed solution allowed us to find 4 optimal natural divisions in the donors' characteristics, both from a customer value perspective and a socio-economic analysis. PVA's marketing department can now start to implement campaigns inspired by the guidelines from the last section.

To further conduct this research it is recommended that the several types of donation solutions represented on PVA's website are analysed and that an affinity analysis is conducted to best find the optimal groups of donations. After understanding the association between the donations, the beginning of extensive research on these combinations can begin, to start up-selling and cross-selling PVA's fundraiser solutions.

REFERENCES

[1] (n.d.), "Assessing clustering tendency." https://www.datanovia.com/en/lessons/assessing-clustering-tendency/.

[2] Wikipedia, "Hopkins statistic." https://en.wikipedia.org/w/index.php?title=Hopkins_statistic&oldid=996091913.

[3] Analytics, B., "How rfm analysis boosts sales." https://www.blastanalytics.com/blog/rfm-analysis-boosts-sales.

[4] Wikipedia, "List of u.s. state and territory abbreviations.." https://en.wikipedia.org/w/index.php?title=List_of_U.S._state_and_territory_abbreviations&oldid=996496793, (Retrieved 1 December 2020).

[5] (n.d.), "Rfm analysis tutorial." https://kaggle.com/regivm/rfm-analysis-tutorial, (Retrieved 29 December 2020).

[6] (n.d.), "Rfm clustering of customers using k-means." https://kaggle.com/jnikhilsai/rfm-clustering-of-customers-using-k-means, (Retrieved 20 December 2020).

[7] Stephanie., "Kaiser-meyer-olkin (kmo) test for sampling adequacy.." https://www.statisticshowto.com/kaiser-meyer-olkin/.